DEVELOPING SCHEDULING STANDARDS USING REGRESSION ANALYSIS: AN APPLICATION GUIDE

Robert J. Graves, PhD
Department of Industrial Engineering and Operations Research
University of Massachusetts
Amherst, MA

Leon F. McGinnis, PhD, PE
School of Industrial and Systems Engineering
Geogia Institute of Technology
Atlanta, GA

June 30, 1987

**Prepared for Robinson-Page-McDonough and Associates, Inc.
as part of Task EC-21 for Panel SP-8 of
The Ship Production Committee**

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
30 JUN 1987	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Developing Scheduling Standards Using Regression Analysis: An Application Guide	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Surface Warfare Center CD Code 2230 - Design Integration Tools Building 192 Room 128 9500 MacArthur Blvd Bethesda, MD 20817-5700**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**49**	

EXECUTIVE SUMMARY

This Application Guide presents a step-by-step introduction to the development of scheduling standards using regression analysis. The presentation employs an example taken from a shipyard sheet metal shop and discusses the issues and procedures in constructing scheduling standards from work order-level data on actual fabrication times.

The methods described in this **Application Guide** have been applied in three different shipyard shops, and in each case have produced scheduling standards with a prediction accuracy of at least 10%, when applied to a set of work orders representing roughly a manweek of work. The cost to establish scheduling standards using these methods compares very favorably to the cost for other techniques, especially if engineered labor standards or measured labor standards must be available for those other methods.

Table of Contents

1. INTRODUCTION

In ship production, as in all other forms of manufacturing, getting the most out of production resources requires knowing how much of each resource, particularly direct labor, will be required for each production job. In the past, estimates for the critical resources, such as the total direct labor hours or the elapsed time to fabricate, were developed manually, using work order data and experience. For a number of reasons, these estimates, regardless of their accuracy for the ship as whole, tend to have substantial errors at the work package level. AS a result, they have not been very useful for purposes such as shop loading.

This Application Guide is the result of research sponsored by the Ship Production Committee through Panel SP-8 into methods for establishing scheduling standards. A scheduling standard is related to traditional engineered work standards, but is designed to indicate shop manhours for relatively large aggregations of work, rather than for use in evaluating alternative work methods.

Because traditional manual methods for estimating the lmr content of work packages are not very accurate, Panel SP-8 has investigated several more formal methods. This Application Guide describes one such method, namely, regression analysis of historical performance data to develop equations for predicting the direct labor content of fabrication operations.

l-l The Intended Use and Audience

This Application Guide is a step-by-step introduction to scheduling standard development using regression analysis. The intended audience is shipyard personnel interested in applying the

statistical approach to scheduling standards as developed under Task EC-21. Although little or no background in statistics is presumed, the use of some technical terms is unavoidable. In some cases, a term or technique will be discussed in the Applications Guide. In other cases, however, the reader is directed to the representative sample of technical references listed in the bibliography.

Several pilot studies (SP-8 tasks EC-13 and EC-21) have amply demonstrated the value of regression analysis as a tool for developing scheduling Standards. Nevertheless, as with any powerful tool, considerable knowledge and judgement are required for its use. An applications team should include someone with formal training in the methods of regression analysis.

To clarify the approach, a case study is presented. This case study is based on an actual shipyard sheetmetal shop, but the data have been coded so that the values described in the report are not actual values. However, the steps described were actually applied to these data, and the scheduling standards generated by regression analysis proved effective for shop loading. Section 2 describes the case study in more detail.

Because there is a variety of software products for regression analysis, this Guide does not cover the details of software usage. Instead, it focuses on the “what and why" of each step in the standads construction process. Thus, while the examples in the guide are based on the SPSS/FC software, the prospective user of this Guide is free to use any comparable system. The authors of this Guide do not endorse SPSS/PC or any other software product. At the time the Guide is being written, there are several other software products that could be used to

perform the regression analysis. The reader is advised to consult an appropriate expert or a software vendor to identify the best alternatives.

1.2 What Is Regression Analysis?

Regression analysis is a technique used to construct a mathematical equation to "explain" observed data. For example, suppose you observe the fabrication of a large number of sheet metal parts, having the shapes shown in Figure 1. For each piece, you record the material, the dimensions of the of the type of seam, the length, and the total direct hours. Regression analysis would allow you to construct a mathematical equation for predicting the direct hours for any piece that was similiar to the ones in your sample of observations.

The direct hours is a response variable, and the attributes (i.e., opening dimensions, length, etc.) are called predictor variables. In developing a regression model to predict direct hours, you must decide which predictor variables to use, and what functional form to use. There usually will be many alternatives, and the objective of the regression analysis is to find the alternative that gives the best predidions with the fewest variables. A simple regression equation might be:

$$TIME = 12.5 + 8.13\ (LENGTH)$$

which says that the time to fabricate a piece of this shape is 12.5 minutes plus an additional 8.13 minutes per foot of length. The numbers 12.5 and 8.13 are called parameters (sometimes coefficients) of the equation and their specific values are determined as part of the regression analysis.

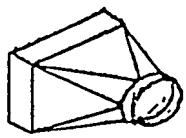
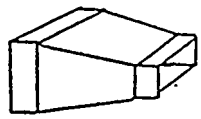
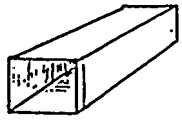
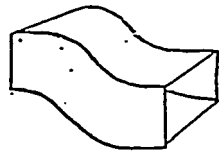
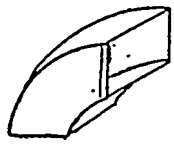
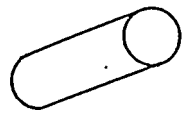
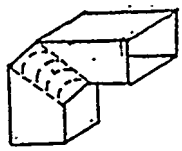
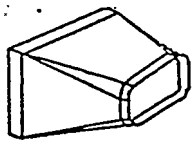
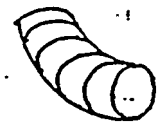
SHAPE No.	SHAPE	SHAPE No.	SHAPE
1		6	
2		7	
3		8	
4		9	
5		10	

Figure 1. Sheet Metal Shapes

More complex forms of the predictor equation are possible. For example,

$$\text{TIME} = \begin{cases} 15 & \text{IF GAUGE = 16} \\ 60 & \text{IF GAUGE = 18, 20, or 22} \\ 40 & \text{IF GAUGE = 24 or 26} \end{cases}$$

indicates that fabrication time is a constant, but different constants apply for different gauges of material. Another form of the predictor equations uses more than one predictor variable:

$$\text{TIME} = 3.13 + 2.56\ (\text{LENGTH}) + 1.87\ (X1*Y1)$$

where Xl and Y1 are the dimensions of a rectangular opening.

Selecting the appropriate predictor variables and developing a good regression model require a combination of: understanding the manufacturing process, understanding certatin aspects of the regression method, and experimentation with alternative models. Equally critical, however, is the collection of accurate data.

1.3 The General Approach

The statistical approach to scheduling standards involves six steps, as illustrated in Figure 2. Product analysis identifies the attributes of the product that may have a significant influence on the direct hours for fabrication. In essence, product analysis determines what data need to be collected.

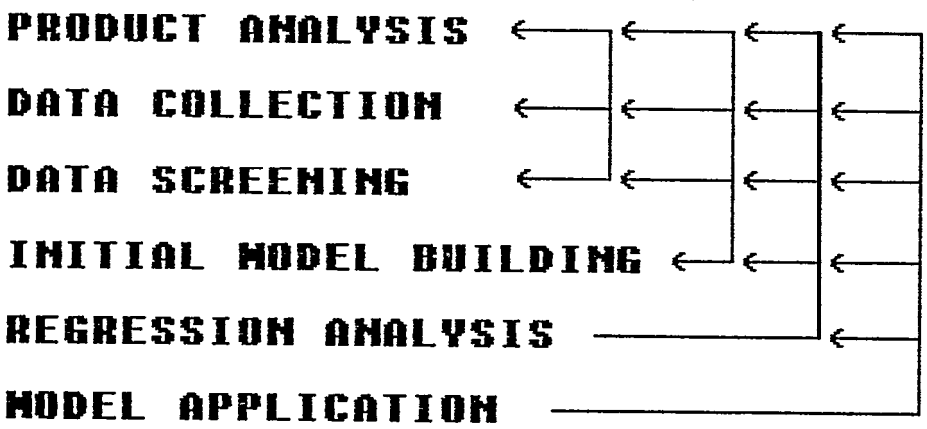


Figure 2. Statistical Approach to Scheduling Standards

Data collection involves selecting the most appropriate method for capturing the actual time and the attributes of the parts being fabricated There usually are several alternative methods for collecting any of the data elements. For example, the actual times may be recorded by the mechanics themselves, or they may be recorded by the supervisor. They might even be taken directly from a material control data base, for instance, if bar codes and readers at each workstation are used.

Data screening is a step in which the raw data from the data collection step is analyzed for gross errors. Examples of the kinds of errors that screening may detect include: incorrect data recording, e.g., keystroke error; missing data; unexpectd data; and invalid data values. Data screening may lead to additional data collection, or to revisions to the product attributes being recorded.

Initial model building involves the selection of the variables to be included in the prediction equation and, perhaps, specification of the form of the equation There are a number of technical analyses that are required in this step to insure the validity of the final prediction equation. Note that these analyses may lead to a reiteration of the product analysis, data collection, and/or data screening steps.

Regression analysis is the step in which the model is refined and its parameters are estimated, based on the data collected in an earlier step. Again, this step involves a number of technical analyses to determine the best set of predictor variables to use, and to insure the accuracy and validity of the predictor equation. Also as before, the analyses may cause a reiteration of some or all of the previous steps.

Application of the predictor equation provides estimates of the time required for a work order. In this step it is important to recognize that there will always be some amount of error in the estimates (“statistical error" is a term often used in this context), and the key is to detect when the errors in the estimates are greater than should be expected.

The remainder of the Application Guide focuses on the analysis steps-data screening, initial model building, and regression analysis. The other steps are discussed in somewhat more detail in the sources listed in the Bibliography.

2. CASE STUDY

An important feature of this application guide is its emphasis on the analysis of a real data set which is specifically oriented toward shipbuilding. For simplicity, throughout the guide, only one data set is used to illustrate the procedures being discussed. The example data set was developed in a study of a sheet metal shop.

Approximately twenty (20) different shapes were produced in the shop. Figure 1 illustrates the most frequently produced shapes. The shapes were analyzed to determine the attributes that were most likely to be important in determining fabrication time. As a result of the analysis, sixteen (16) attributes were defined, and are listed in Appendix A. Not all of the sixteen attributes are relevant for every shape, e.g., ANG does not apply to straight shapes, such as shape number six. However, together, the sixteen attributes are sufficient to describe any of the shapes.

Data collection in the sheet metal shop involved two activities. First, the mechanics were instructed to record the actual fabrication time for each detail in a sample of work orders. Second, the attributes of each of the details were obtained from the work packages for the sample of work orders. The two sets of data then were combined to obtain the regression analysis database. Appendix B presents a listing of the database for shape 1 only, where TIME values have been coded to disguise the actual times. Only shape 1 is discussed in this manual. Whenever "the data set" is discussed in the following sections, the reference is to the data set listed in Appendix B.

3. PRELIMINARY DATA SCREENING

Screening data prior to formal regression analysis is critical. Even one or two "bad" data values can significantly change the final regression equation and lead to poor predictions, especially when the size of the data set is small. "Bad" data in a sample can result from any one of a number of factors. There may be errors in recording or transcribing data. The "rules" for recording the data (e.g., what time is to be included) may not be clear to the recorder (often the operator/mechanic) or may not be followed. Breakdown of key equipment or other unusual events can distort the data.

Bad data also will almost surely result if there is an adversarial relationship between the "standards" people and the "production" people. The view held by the EC-21 project team is that regression-based standards are not a tool for making people work harder. Rather, they are a tool for making management smarter about loading work onto facilities and people so that the schedule represents a realistic worlload.

There are several steps that should be taken to screen the data sample. While these steps will not guarantee a “good” sample for subsequent analysis, they should eliminate most of the gross errors. The following-sections discuss these steps.

3.1 Scanning the Data Visually

Some errors are bound to occur when recording or transcribing data. One of the ways to spot such errors is simply to look through a listing of the data. Extremely large or small values, incorrectly typed decimal places, out-of-place minus signs, missing values, etc., are a few examples of common errors.

Several examples of data recording errors can be seen in the data set. six records, 18, 89, 90, 179, 200, and 282, appear to be highly questionable because they contain unexpected information. Since shape 1 has only one rectangular opening, measurements X2 and Y2 should not be present. In several records, there are missing values. Records 67, 68, 91, 129, 140, 176, and 213 are missing material codes.

Regardless of the cause of the error, the analyst must deal with it. The best response would be to go back to the original records to obtain the correct information. Unfortunately, it often is not possible and the only option left to the analyst is to decide whether or not to eliminate the records from further analysis.

In the above example, all six records containing extra information were eliminated from the data set, but the records with missing values were not. The particular missing values may not be needed in the regression model, so there remains, a possibility that the records can be used.

Probably the most difficult errors to deal with are extreme values, or "outliers". Potential outliers are extremely large or small observations which are not typical of the remaining data points. For example, in record 91, TIME has a relatively high value and needs to be examined If the outlier is a result of recording error, it should be either corrected or eliminated from the database immediately. But there may be cases where the unusual value may explain an interesting aspect of the process. In such cases, eliminating that value arbitrarily might result in a limited prediction equation. In general, the analyst would want to have an assignable cause for dismissing a data point as an outlier. Therefore extreme values should not routinely be considered an outlier and deleted from analysis.

In the example data base, the large value of TIME cannot be described as an error in the data recording, because the record seems to be consistent and complete. Also, the value is permissible according to the definition of the variable TIME. In the absence of some specific reason arising from the fabrication process under study, the value should not be deleted from the data set, since the large value might be due to the large values of X1 and Y1. This particular data point will be discussed more in Section 3.3.

3.2 Descriptive Statistics

Descriptive statistics include frequencies, means, standard deviations, maximum and minimum values, etc., and provide valuable information about each variable Means tables and crosstabulations provide insight into the data. Using such methods is especially important when the database is large, since it is easier to read the summary statistics than a listing of all of the records. Another

purpose for obtaining summary statistics is to become familiar with the data. Developing an intuitive understanding of the relationships among the variables enables the analyst to define the initial model properly.

To appreciate the usefulness of descriptive statistics the following example is given. The table that follows is taken from an SPSS/PC output and displays the frequencies for MATL in the example data set.

Table 1. Material Frequencies

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
GALVANIZED STEEL	1	15	37.5	45.5	45.5
STAINLESS STEEL	3	18	45.0	54.5	100.0
		7	17.5	MISSING	
	TOTAL	40	100.0	100.0	

Valid Cases 33 Missing Cases 7

Examination of a listing of the records, sorted by type of material, reveals that it is not worthwhile to make MATL a predictor variable in the analysis for two reasons. First, material type is missing for 17.5% of the cases, which is excessive. Second, one category, perforated aluminum, is not represented at all. This would limit the model to estimating only certain material types.

Another useful tool to screen a categorical variable is the means table for each category. The following table is obtained from an SPSS/PC output and displays the average times for each different seam type.

Table 2. Mean TIME by SEAM Type

Summaries of TIME
By levels of SEAM

Variable	Value	Label	Mean	Std Dev	Cases
For Entire Population			77.6538	106.1325	39
SEAM	2	RIVET	132.9167	171.6426	12
SEAM	4	WELD	45.5000	27.2534	3
SEAM	5	3/4" LAP	20.0000	0.0	1
SEAM	6	SPOT WELD	44.8000	20.4353	15
SEAM	8	LAP	75.6250	73.0307	8

Total Cases = 40
Missing cases = 1 OR 2.5 PCT.

Note the significant differences in average times for different seam types. This indicates that different predictor equations may be required for different seam types, or SEAM will have to be included as a variable in the predictor equation. But, as with material, there is not enough data for each seam type to include SEAM in the prediction equation. In particular, seam types 4, 5 and 8 are not adequately represented in the sample.

The means table clearly indicates that the sample data is not uniformly distributed across seam types. When there are a number of potential predictor variables, the distribution of the sample data across these predictor variables is extremely critical. If the sample has predictor variable values only over a small part of the possible range of values for that predictor variable, then it is quite possible that the regression model will be used to extrapolate, i.e, predict fabrication times for shapes that are outside the range of the original modelling database. Extrapolation should be avoided because of the risk of significant prediction errors.

Crosstabulations provide a convenient mechanism for establishing the boundaries for applying the regression model, and also for determining where additional data may need to be collected Table 3 illustrates a crosstabulation for shape 1, showing SEAM by MATL.

Table 3. Crossbhlation of SEAM with MATL

Crosstabulation: SEAM By MATL

MATL-> / SEAM	Count	1	3	Row Total
	2	11		11
	4		3	3
	5	1		1
	6		15	15
	8	2		2
	Column Total	14	18	32
		43.8	56.3	100.0

Number of Missing Observations = 8

The conclusion to draw from Table 3 is that it would not be proper to develop a regression model with MATL and SEAM as predictor variables. For each material type, there really are only enough records for one seam type, but it is a different type for each material.. If it were possible to select the shapes to be included in the sample, then shapes with material type 1 and seam types 5 and 8 should be emphasized, along with material type 2 and seam type 4, to achieve a better balance in the data.

3.3 Scatterplots

Scatterplots, or two variable plots are an important part of every regression analysis and should be examined routinely prior to formal analysis. Scatterplots are indispensable devices to detect possible outliers as well as to investigate the relationships among variables. But as pointed out earlier, plots should not be used to eliminate a data point just because it is different from the remaining data points.

The plot of TIME vs. Y1 shown in Figure 3 was obtained from an SPSS/PC output.

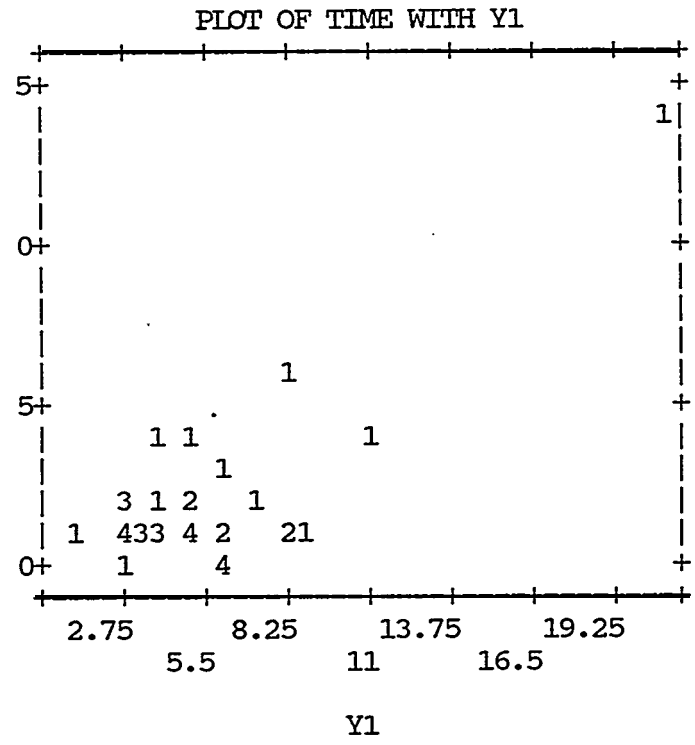


Figure 3

Careful examination of Figure 3 suggests that the observed relationship between TIME and Y1 might be at least partially explained by a linear equation through the origin. In Figure 4, such an equation has been placed on the scatterplot so that it “appears” to be as close as possible to “most" of the points. Obviously, there is no straight line

that covers all of the points in the scatterplot. How well any one straight line "explains the data" depends on how close the observed values are to the line, or how much error there is in representing them using the line. In this example, most of the observations fall close to the line, but a number of them are spread away from the line.

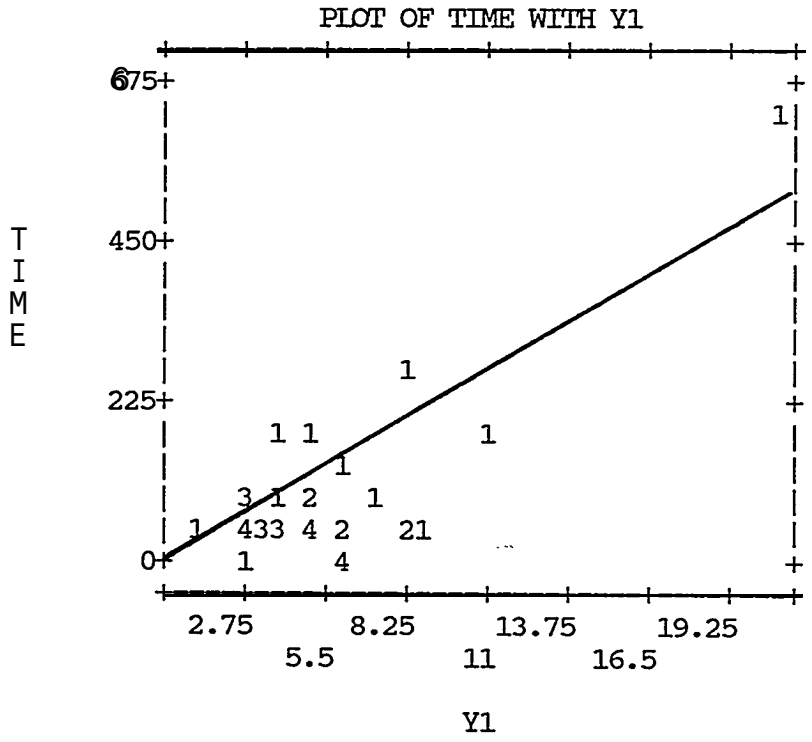


Figure 4

The scatterplot in Figure 3 also shows that, with one exception, the points are all grouped together. The one exception is the point in the upper right hand corner of the scatterplot. This point might seem to be an outlier, because it appears to be so different from the other points. Note, however, that it falls reasonably close to the straight line in Figure 4. Furthermore, it cannot he described as an error in collecting or entering the data. This unusual observation is simply due to the large values of Xl and Y1 and the record should not be eliminated from the database.

4. INITIAL MODEL BUILDING

Once the database has been screened and any obvious errors or problems have been dealt with, the next step is initial model building. In this phase of the statistical approach, the data is analyzed in several different ways, with the objective of discovering which attributes should be used in the predictor equation, and identifying how the attributes should contribute to the prediction of actual time.

Initial model building is a crucial part of regression analysis. Effort is usually required in setting up the data base to enable analysis with the software and in determining which variables should be included in the initial model. Similarly, as analysis proceeds, the analyst may need to transform predictor variables and should be certain that they are defined in proper functional form. Problems which are most likely to arise when defining variable types, selecting variables, and building a model are discussed in the following subsections.

4.1 Numerical and Non-numerical Variables

In the examples cited in section 3, several types of variables were present. One type of variable is a continuous numerical variable, such as the dimensions Xl or Yl, or the length of a sheet metal shape. Another type of variable is not numerical, such as material type or seam type, and reflects a category or qualitative feature of the workpiece. Both of these variable types are potentially important as predictor variables and, through proper handling, may be included in the analysis.

4.1.1 Numerical Variables

Numerical variables can be handled in a straightforward manner. A prediction equation given by

$$TIME = 4.09 + 2.3(X1) + 2.2(Y1)$$

means that as the variable Xl increases, it will have the effect of increasing TIME by a factor of 2.3 for each unit of increase in X1. Similarly, the variable X2, if increasing, will increase TIME by 2.2 for each unit increase. In like manner, a decreasing variable value will decrease TIME. For different workplaces with different dimensions Xl and Yl, the prediction equation is used to calculate TIME.

Therefore, the numerical-valued variables can be directly accommodated in the model building process. For the non-numerical variables, additional steps must be taken before proceeding further toward model building.

4.1.2 Categorical Predictor Variables

Nonnumerical observations are described as categorical or qualitative variables. For example, in the data set, material type is a qualitative variable with three categories; galvanized steel, perforated aluminum, and stainless steel. In regression analysis, all categorical variables must be given numerical codes for the analyses performed by the software. For instance, codes could he assigned as: galvanized steel = 1, perforated aluminum = 2, stainless steel = 3.

These numerical codes often are chosen for efficiency in data collection However such a choice for coding for analysis purposes should be avoided since it might mask the true influence of the categorical variable on the response. By assigning values 1, 2, and 3 we are forcing each material type to have a precise effect on the

predicted response. If aluminum actually has the most or the least effect on time, the code assignment given above cannot reflect this.

Then, what would be the proper way of coding the categorical variable? In the example above, the three material categories are efficiently coded for data collection purposes using only one variable, which can assume three clifferent values:

$$X1 = \begin{cases} 1 \text{ if galvanized steel} \\ 2 \text{ if perforated aluminum} \\ 3 \text{ if stainless steel} \end{cases}$$

Note that with this type of coding, there is an implied ordering of the categories, and a constant "contribution" to the response. In other words, the contribution of perforated aluminum is twice that of galvanized steel, and the contribution of stainless steel is three times that of galvanized steel. In reality, there may be no such ordering, and the contributions may be nonlinear.

The correct way to cede for material for the analysis in this example, would be to use two indicator variables. Let

$$X1 = \begin{cases} 1 \text{ if galvanized steel} \\ 0 \text{ otherwise} \end{cases} \qquad X2 = \begin{cases} 1 \text{ if perforatedaluminum} \\ 0 \text{ otherwise} \end{cases}$$

Note that X1=X2=0 implies that the material is stainless steel. Using indicator variables to code material implies no prior ordering of the categories. Furthermore, it allows the contribution for each material type to be uniquely defined, instead of defined relative to the other categories. In general, if a non-numerical attribute has k different categories, then k-1 indicator variables should be defined to code it.

An additional benefit of using indicator variables to code categories is that if there is not enough data in the database to support accurate model building for a particular category, then the

resulting model would isolate this fact by not including the corresponding indicator variable.

4.2 Selecting Variables

The prior section addressed several issues regarding variables as a first step to prepare the raw data for subsequent analysis. Another step is one that also occurs prior to analysis when the nature of predictor variables may be altered in order to improve prediction of the response variable. This required when a nonlinear relationship between predictor variables and the response variable appears to exist.

One type of alteration is discussed, i.e., creating new predictor variables from the products of two or more individual predictor variables, along with the means to identify when it is necessary or desirable to consider doing so. Several other types of alterations are not discussed, e.g., adding powers of individual predictor variables or logarithmic transformations, since these are considered beyond the scope of this Guide. Those readers interested in discussions of more elaborate variable transformations are referred to sources in the bibliography.

4.2.1 Transformations

One useful transformation that should be mentioned is the addition of “interaction terms”. Interaction terms in a regression model are products of two or more predictor variables. They are useful when it is believed that the predictor variables have a joint influence, as distinguished from individual influence, on the response variable. Several interaction terms created by using two or more predictor variables can be included in the model, but they should not be inserted

routinely for two reasons: first, the number of possible interaction terms can be very large and might result in unnecessarily complicated models; second, adding interaction terms might repeat the information provided by the individual predictor variables and result in multicollinearity, a problem which is discussed in the following section.

Routinely plotting the data not only helps to detect errors and outliers, but also aids in the examination of relationships among the variables as well. Scatterplots of response variable vs. predictor variables can be used to investigate the functional relationships between the dependent variable and each predictor variable individually. In some cases, the plot might suggest a nonlinear relationship which requires a transformation of that predictor variable.

In our case study it is reasonable to believe that the measurements Xl and Y1 may have a joint influence on TIME, since 2 (Xl+Yl) is the circumference and X1Yl is the area of the corresponding opening. Therefore, it is worthwhile examining the influence of a new variable, say X1Y1, which is created by multiplying Xl and Y1 as an addition to the model. Following is a plot of TIME vs. X1Y1 obtained from an SPSS/PC output.

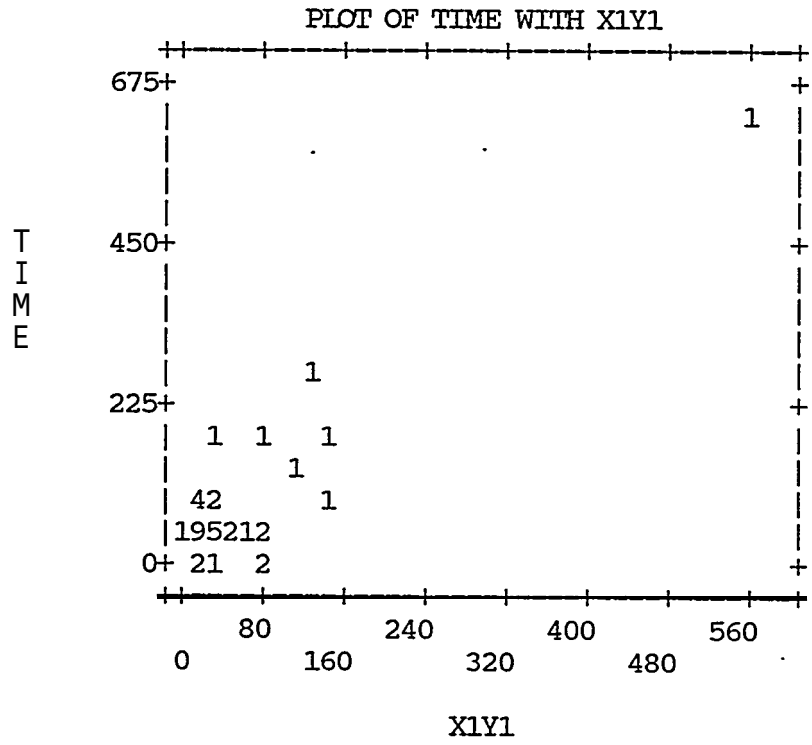


Figure 5. TIME vs X1Y1

Careful examination of this plot suggests that the interaction term, X1Yl, might be a valuable addition to the specification of the regression model. Comparison of this plot with the Figure 3 plot of TIME vs. Y1 alone reveals that the data is better described by a linear relationship between TIME and X1Y1.

4.2.2 Multicollinearity

Selecting predictor variables for use in a prediction equation, when there are several variables available, is a required task in model. building. An important goal in this variable selection process is to include the most influential predictor variables in the final prediction equation. However, the process of selecting variables must consider the correlations among the predictor variables. In some cases, one predictor variable may repeat the information provided by another predictor variable. This type of redundancy is referred to as a "multicollinearity". Since strong correlations among the predictor variables will likely result in poor prediction equations, it is important to know how to detect and deal with multicollinearities.

Pairwise correlations between predictor variables can be identified from the off-diagonal elements of a matrix of correlation coefficients. This is a standard analysis available in any well designed regression analysis system. As a rule of thumb, any pairwise correlation larger than 0.70 or 0.80 indicates that at most one of the pair of predictor variables should be included in the model.

A correlation matrix of the predictor variables being considered for inclusion in the regression model for shape 1 is presented in Table 5. Only 34 records are represented in the correlation analysis, due to missing data. Several things should be noted about this matrix. First, all the entries on the main diagonal equal one. This is because these entries show the correlations of the variables with themselves, and every variable is perfectly correlated with itself. Also, the entries in the correlation matrix are symmetrical around the main diagonal.

Table 5. Correlation Matrix

Coefficients have been calculated through the Origin.

N of Cases = .34

Correlation:

	TIME	GAUGE	Xl	Y1	DIAM	XlYl
T I M E	1.000	.522	.782	.830	.791	.942
GAUGE	.522	1.000	.787	.798	.802	.437
Xl.	.782	.787	1.000	.894	.911	.805
Y1	.830	.798	.894	1.000	.945	.836
DIAM	.791	.802	.911	.945	1.000	.768
XlYl	.942	.437	.805	.836	.768	1.000

Examination of the first row of the correlation matrix reveals that TIME is highly correlated with the predictor variables X1Yl, Yl, X1, and DIAM, since the correlation coefficients all are close to one. On the other hand, GAUGE has a fairly low correlation with TIME, indicated by a correlation coefficient value of only 0.522. The of pairwise correlations between the predictor variables, X1, Y1, DIAM, and X1Y1, all are greater than 0.75, which suggests that these variables are highly interrelated. Because of the large correlations, it would not be wise to include more than one of the predictor variables X1, Y1, DIAM, or X1Y1 in the regression model.

4.3 Initial Model Building

Sometimes the scatterplots suggest that the relationship between the response and the predictor variables is linear through the origin. This situation raises the question of whether an intercept parameter should be included in the model specification. The first point to note about no-intercept models is that they require the response to be zero when the predictor variable is zero.

When using SPSS/FC, the type of initial model may be specified as through the origin or not through the origin. If a specification of not through the origin is made, the analyst is assuming that some characteristic of the physical process, e.g., setup time, will be present in the data and should be reflected in the model.

For the case of shape 1, the correlation analysis and scatterplots suggest that a good initial model is:

$$TIME = b1 * (X1Y1)$$

This model and its variations must be analyzed to determine how well it fits the data.

Having now studied the types of variables to be defined in the data, the types of variables to be considered for inclusion in the model, as well as the general initial model form, it is appropriate to consider the details of model development through regression analysis. This subject is discussed in the next section.

5. ANALYSIS

In this section, the use of “the least squares parameter estimation method" to fit a regression model will be discussed. Although computers will be used to execute the analysis, it is important for the user to understand the operations and statistics that are used in the analysis. After all, computers are not capable of making the necessary interpretations and judgments that lead to building the correct model.

The starting point for this phase of the analysis is the database, which has been screened to correct obvious errors, and the preliminary selection of the predictor variables and the mathematical form to be used for the prediction equation. This phase of the analysis accomplishes several critical objectives:

1. The mathematical form of the predictor equation is "fitted" to the data, i.e., particular values are computed for the coefficients of the predictor variables.
2. The "goodness of fit" is evaluated for the resulting prediction equation.
3. The fitted model is evaluated to determine if there are any datapoints which are (statistical) outliers and should be eliminated from the dataset.

Each of these three elements of the analysis depends on the mathematical theory of regression analysis. However, there is no cookbook method for using the theory that will automatically lead to the correct result for the analysis. Knowledge of the theory must be combined with judgement and an understanding of the manufacturing process to insure a good result.

5.1 Fitting The Model

The result of preliminary model building is a mathematical form for the prediction equation where the values of the coefficients are, as yet, unknown "Fitting the model" involves determining specific values for these coefficients.

The basic computational procedure for determining the coefficient values is called the least squares regression method. Essentially, what this method does is to determine values for the coefficients so that the differences between the actual values of TIME in the database and the corresponding values from the predictor equation, added up over all records in the database, are as small as possible.

Since some of the errors will be positive and some negative, the errors are squared prior to adding them up, so as not to give an overly optimistic assessment" of the total error. Thus, "least squares" refers to finding the model fit that has the smallest sum of the squared errors.

There are several methods that can be used to evaluate the accuracy of the prediction equation. These methods all attempt to provide an answer to the following question, "How well does the predictor equation explain the observed data?" If the predictor equation does a good job of explaining the observed data on which it is based, then there is some reason to believe it will do a good job of predicting for new observations. The following sections describe the use of some standard statistics for assessing the goodness-of-fit.

5.1.1 Analysis of Variance Table

In statistical methods, the variability in a set of observations, e.g., observed values for TIME, is related to the sum, over the sample, of the differeces between each observation and the mean, or average, for the sample. This quantity often is referred to as the "(total) sum of squares" for the sample, and is constant for a given sample.

After fitting a regression model, the total sum of squares can be partitioned into two portions and presented in a summary table called the "Analysis of Variance Table" (abbreviated as ANOVA table). SPSS/PC provides the following table for a prediction equation of the form:

TIME= B1 * (X1Y1)

Only 38 records are used in the analysis, since two records are missing values for X1 and Y1.

Table 6. ANOVA Table

Variable(s) Entered on Step Number
1 X1Y1

Multiple R	.92556
R Square	.85667
Adjusted R Square	.85279
Standard Error	50.81858

Analysis of Variance

	OF	sum of Squares	Mean Square
Regression	1	571106.19325	571106.19325
Residual	37	95553.55675	2582.52856

F = 221.14226 Signif F = .0000

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
X1Y1	1.14755	.07717	.92556	14.871	.0000

The total sum of squares is divided into two components, called the regression sum of squares (571106.19325) and the residual sum of squares (95553.55675). The regression sum of squares corresponds to the differences between the predicted values and the sample average, and is a measure of the sample variability that is explained by the regression. The residual sum of squares measures the variability in the sample that is left unexplained after considering the regression model.

The column in the ANOVA table labeled DF is the "degrees of freedom," an indication of how many contributors there are for the sum of squares. The degrees of freedom for the regression sum of squares is the number of independent variables, say k (in this case k=1), and for a through-the-origin model the degrees of freedom for the residual sum of squares is n - k (in this case 38-1=37), where n is the number of records in the sample database. The mean square figures can be obtained by dividing the corresponding sum of squares by their degrees

of freedom. These mean squares figures are then used to test the overall significance of the regression relationship by computing the F ratio.

The F ratio is a statistic defined as:

$$F = \frac{\text{Mean square due to regression}}{\text{Residual mean square}} = \frac{571106,19325}{2582.52856} = 221.14226$$

If this ratio is large, it indicates that the amount of variation in the response variable that is explained by the predictor equation is large relative to the amount of variation that is left unexplained by the predictor equation. Therefore, the larger the F ratio is, the more acceptable the model is. Testing the significance of the ratio involves using standard tables of the F statistic. The SPSS output displayed in Table 5 is interpreted as follows: there is less than a 1 in 100,000 chance of getting an F value of 221.14226 "by accident." Statistically at least, this is a very strong indication of a good fit.

5.1.2 Error Variance

Another statistic that is used as a measure of accuracy is the error variance which can be interpreted as the unexplained variability of the responses. In the computer output given in Table 6, it is designated as "Standard Error" and represents the standard deviation of the observed time around the regression line at the average observation. For the example output, the standard error is 50.81858, indicating that approximately two-thirds of the times that shape 1 fabrication time is observed, it will lie within ±50 minutes of the regression equation's predicted values. The smaller the error variance, the more accurate the fit.

5.1.3 Coefficient of Determination

The coefficient of determination, labeled "R Square" in the output is used to measure the strength of the relationship between the response and the regression equation. It is one of the most important measures of model adequacy. R Square gives us the percentage of the variability of responses that can be explained by the regression equation.

In the output shown in Table 6, Multiple R is the square root of R Square. The figure labeled "Adjusted R Square" modifies R Square since it tends to be a too optimistic figure. If the residuals are small, R Square will be close to one; but if they are large after the regression model is fitted, R Square will be close to zero. However, a large value of R Square alone does not necessarily guarantee an accurate prediction The error variance also should be considered.

5.2 Variable Selection Methods

There are times that the analyst, constructing a multiple regression equation may have a set of fifteen or twenty independent variables that might be associated with the response variable of interest. However, it would be impractical to construct a regression equation using all fifteen or twenty variables. In this case, the analyst may wish to construct a regression equation using only a subset of the original. predictor variables in order to obtain an equation of a manageable size. However, this raises the following question, "which subset of the predictor variables would be the most influential on the prediction equation?" A procedure that often is used in this situation is "stepwise multiple regression."

The stepwise multiple regression method selects the best set of predictor variables. At each step two decisions are made: which variable (if any) to enter into the model; and which variable (if any) to remove from the model. These decisions are made using several criteria, in such a way that a variable enters the model only if it improves the statistical measures of fit, and is removal from the model when doing so does not degrade the statistical measures of fit. The procedure halts when a step neither enters nor removes a variable.

For shape 1, a stepwise regression analysis was performed, using the predictor variables GAUGE, X1, Y1, DIAM, and X1Y1. Table 7 presents the output for the first step, in which X1Y1 was entered into the model. Recall that X1Y1 was the predictor variable with the highest correlation with TIME. Note that only 34 records are used in the analysis, due to missing values for some of the predictor variables.

At the end of the first step, the adjusted R square value of .88 is quite good, and the standard error is 43.9, or 60% of the average TIME. The four predictor variables not in the model are evaluated in the section of Table 7 labeled "Variables not in the Equation" A variable with a small value of "Min Toler" or a large value of Sig T" is not a good candidate to enter the model on the next step. Note that X1, Y1, and DIAM are not good candidates to enter, confirming the conclusion reached earlier from the pairwise correlations.

Step 2 of the analysis is summarized in Table 8. GAUGE enters the model, X1Y1 is not removed, and there are no more good candidates to enter, so the procedure ends. Note that the measures of model fit (Adj R Square, Standard Error, F) have not been improved dramatically by adding GAUGE. The rule of parsimony (if two models are more or less equivalent, use the one with the fewest predictor variables) would

dictate using the model with just X1Y1.

Table.7. Stepwise Regression: Step 1

Variable(s) Entered on Step Number
1.. X1Y1

Multiple R .94192
R Square .88721
Adjusted R Square .88379
Standard Error 43.88291

Analysis of Variance

	DF	Sum of squares	Mean Square
Regression	1	499886.34015	499886.34015
Residual	33	63548.40985	1925.70939

F = 259.58555 Signif F= .0000

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
X1Y1	1.11018	.06891	.94192	16.112	.0000

Variables not in the Equation

Variable	Beta In	Partial	Min Toler	T	Sig T
GAUGE	.13689	.36667	.80922	2.229	.0329
Xl	.06845	.12094	.35208	.689	.4957
Y1	.14080	.23030	.30174	1.339	.1901
DIAM	.16429	.31321	.40993	1.866	.0713

Table 8. Stepwise Regression: Step 2

Variable(s) Entered on Step Number
2.. GAUGE

Multiple R	.94993
R Square	.90238
Adjusted R Square	.89627
standard Error	41.45955

Analysis of Variance

	DF	sum of squares	Mean Square
Regression	2	508430.12301	254215.06150
Residual	32	55004.62699	1718.89459

F = 147.89450 Signif F= 0.0

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
X1Y1	1.03971	.07237	.88213	14.367	.0000
GAUGE	.77120	.34591	.13689	2.229	.0329

Variables not in the Equation

Variable	Beta In	Partial	Min Toler	T	Sig T
X1	-.30031	-.33030	.11810	-1.948	.0605
Y1	-.24259	-.20468	.06949	-1.164	.2532
DIAM	.02499	.03008	.14138	.168	.8680

End Block Number 1 PIN = .050 Limits reached.

5.3 Residual Analysis

The term "residuals" refers to the differences between the predicted and observed values of the response variable. Examining residuals is an important tasks in any regression analysis, because it helps to spot any errors that are overlooked during the initial screening of the database as well as to detect any misspecifications in the model form. There are both graphical and numerical methods

available for examining residuals. First the graphical methods are discussed and this is followed by a presentation of an analytical method.

5.3.1 Graphical Methods

Graphical methods employ plots of the residuals with either the predictor variables or the predicted or observed values of the response variable. There are three commonly used plots that are discussed.

Normal Probability Plots

A fundamental assumption in least squares regression is that the residual errors are randomly distributed and follow a normal distribution. One way of checking this assumption is to plot the ordered residuals on a special graph paper called normal probability Paper. If the residuals are normally distributed such plots should approximately follow a straight line through the origin Serious deviations from a straight line may suggest a need to transform some of the predictor variables, or to consider alternate forms for the predictor equation. An individual value far off the line may be an indication of an outlier observation that does not follow the model.

In the example given in Figure 6, the residuals generally follow a straight line, even though there is a tendency to fluctuate away from the line for a few points in the middle. Minor variations and fluctuations of three or four points are common in normal probability plots even when the data is normally distributed. Such fluctuations may be reduced as the sample size is increased.

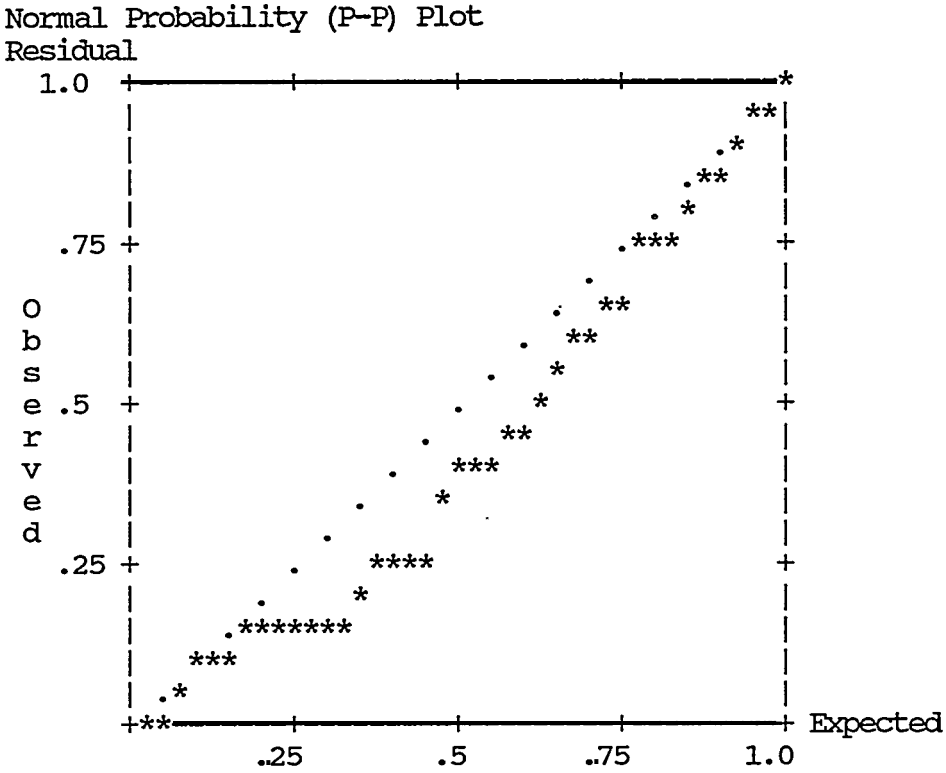


Figure 6. Normal Probability Plot

Plots of Residuals vs. Predicted Values

Another important assumption in least squares regression is that the variance of the residual errors is constant over all the predictions. A simple approach to testing this assumption is to examine plots of the residuals vs. the predicted responses. In this plot, the points should follow a horizontal trend centered at zero, and the spread of the points should be about the same over all the predicted values. If the spread varies, the implication is that the error variance is not constant. If the trend is other than a horizontal line, then We errors are not randomly distributed, indicating an inadequate model. Another advantage of this plot is that it can detect outliers as well as it detects model misspecifications. The plot in Figure 7 was obtained from an SPSS/PC output. Note that both the residuals and the predicted values are standardized

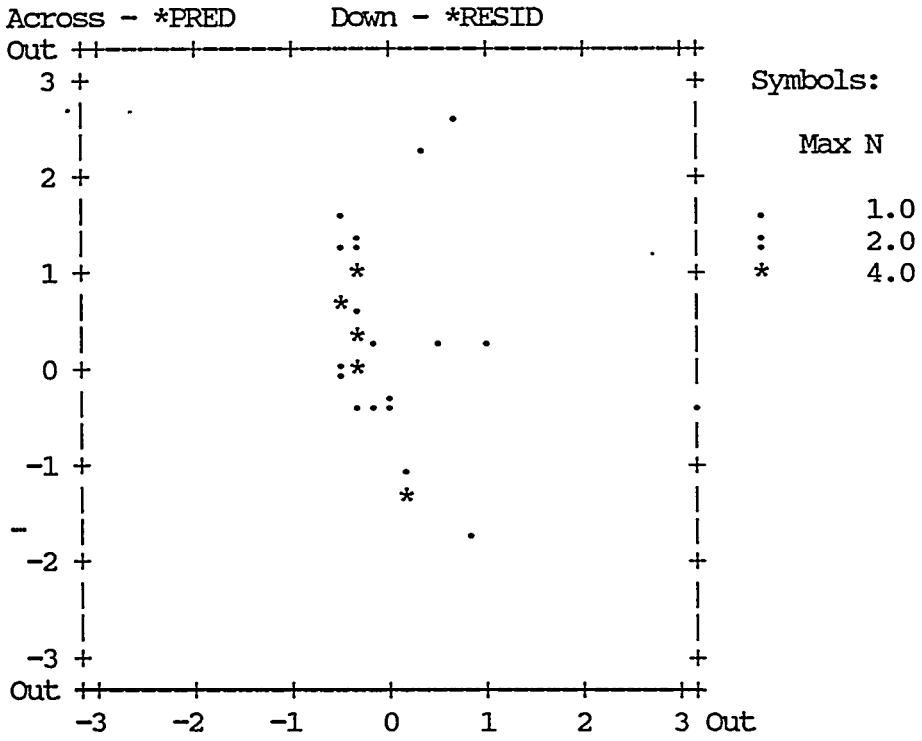


Figure 7. Residuals vs Predictions

Plots of Residuals vs. Predictor Variables

Plots of residuals against predictor variables indicate the distribution of the residuals as a function of the predictor variable. Ideally, the plotted points will follow a horizontal trend line, centered at zero, with uniform spread or dispersion across the predictor variable values. Any other trend would be an indication of a need for a transformation of the predictor variable, or perhaps additional predictor variables in the model. The plot displayed in Figure 8 was obtained from an SPSS/PC output, and shows the residuals with DIAM, a predictor variable not in the model. Note that both the residuals and the predictor variable are standardized Normal.

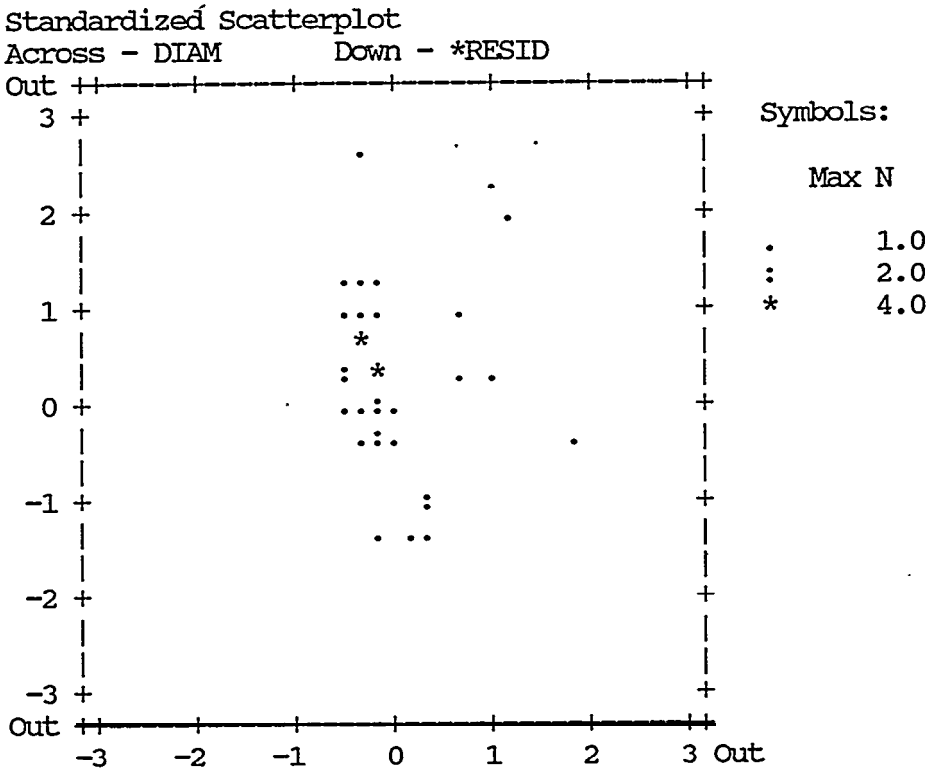


Figure 8. Residuals vs. DIAM

Careful examination of the graph shows no systematic patterns. All the residuals are randomly distributed around the horizontal line. Since this is a graph of residuals against the predictor variable, DIAM, which is not in the model, it indicates that it is not worthwhile to include DIAM in the equation.

5.3.2 Numerical Methods

As mentioned earlier there are also numerical methods available to detect data errors and initial model misspecifications. There are many different types of residuals, although each is a function of the difference between the observed and the predicted responses. There are, for example, raw residuals, standardized residuals, deleted residuals, and studentized residuals. Each of these residuals has different properties and can be useful in different situations.

However, this section contains definitions and discussions of several techniques where the emphasis will be primarily on the interpretations of SPSS/PC outputs.

The example in Table 9 was obtained from an SPSS/PC output for the model TIME = b1*(X1Y1). It is a summary report of residuals to warn the analyst about any potential outliers. The first row of the table gives the summary statistics of the predicted responses, designated as *PRED. These statistics include the minimum and maximum values, average value (mean), standard deviation and sample size respectively. The second row gives the similar statistiCS about the residuals, designated as *RESID. The third and the fourth rows contain the summary statistics about the standardized predicted responses and residuals, designated as *ZPRED and *ZRESID respectively. Standardization is used to allow comparison of specific sample results to standardized tables of "expected" results.

Table 9. Residuals statistics

	MIN	MAX	MEAN	STD DEV	N
*PRED	5.7377	650.6594	63.4322	106.3152	38
*RESID	-70.6566	143.2785	16.5547	47.9694	38
*ZPRED	-.5427	5.5235	-.0000	1.0000	38
*ZRESID	-1.3904	2.8194	.3258	.9439	38

To help spot outliers, SPSS/FC provides a listing of the standardi zed residuals for the 10 cases with the largest residuals in terms of absolute values. Table 10 contains a listing of the 10 worse residuals for our model.

Table 10. **Outlier Analysis**

Outliers -Standardized Residual

Case #	*ZRESID
30	2.81941
3	2.30810
31	2.03068
32	1.46616
39	-1.39037
18	1.26615
27	-1.26181
6	-1.23229
17	1.22906
9	-1.18310

Residuals with large values do not necessarily mean that there are outliers in the data, so records that appears on the list should not be identified automatically as outliers. In SPSS/PC, only the standardized residuals larger than 3 are identified as outliers. Hence, in our example above none of the values are qualified as outliers. This result was expected since none of the residual plots that were examined earlier contained any isolated points above three or below three from the horizontal.

5.4 Summary

Least squares regression is not just a matter of executing a single computational procedure. It involves not only calculating the estimates for the coefficients in the model form, but also careful examination of a number of statistics to insure that the fitted model is accurate and adequate.

Our experience with using regression analysis to develop scheduling standards has been that a considerable amount of experimentation is required. A number of different model forms, variations of model form, transformation of variables and elimination of outliers usually will be

considered before discovering a "best" fitted model. Again, because this is a process of discovery, considerable judgement is exercised in deciding when to stop looking for a better model.

6. USING THE MODEL

The purpose for developing the predictor equation is to be able to establish scheduling standards for future work orders. Because the predictor equation is based on a sample of the work k the shop during a particular period of time, some care is required in its application.

Assuming that the analysis steps described in sections 3, 4 and 5 have been completed, a regression model that adequately describes the sample of work has been developed. Predictions from this regression model will provide a sound basis for shop loading, provided the work mix being estimated is sufficiently similar to the work mix in the sample from which the model was developed.

This requirement for similarity of work mix leads to a two-phase approach to application. In phase one, the prediction equation is tested to insure that, even though the model is a good one statistically, it is also adequate for shop loading. In phase 2, the predictor equation is used to load the shop for ongoing operations. These two phases are discussed briefly in the following sections.

It is important.ant to bear in mind that there will always be some error between the estimated direct hours and the actual direct hours for a given work order. If the predictor equation is valid, over a large sample of work orders, these errors will add up to a small sum, since some of them will be positive and others will be negative.

6.1 Testing the Model

Model testing insures that the model which is statistically "best" is also accurate enough for use in loading the shop. Testing consists simply of predicting the workload for a sample of work orders, then comparing the predictions to the actual direct hours. The sample of work orders used to test the model must be different from the sample used to construct the model, and must be large enough so that statistical errors in the individual item estimates are not mistaken for misspecification errors.

There are two ways in which the predictor equation may “fail" this test (assuming that the model was evaluated as having a “good” fit in the regression analysis). In both cases, the indication of failure is that the total predicted workload is quite different from the total actual direct hours.

One type of failure occurs when the work mix changes from the sample used to develop the prediction equation to the sample used to test the model. The conclusion would be that the original sample was not “representative” of the total work mix in the shop, and therefore was inadequate for developing a scheduling standard prediction equation. The resolution of the problem would involve expanding the sample database and reiterating the regression analysis.

This type of failure is due to a common error in applying regression models that is called ‘extrapolation’. The regression equation constructed from a particular data set is valid for making predictions of the response variables, only if the independent variables used in the predictions are within the range of the original set. For example, in the case study analyzed here, the measurement Xl used to construct the regression equation ranged from 2 inches to 27 inches.

Therefore, attempting to predict time for a sheetmetal piece with an opening height of 30 inches would be extrapolating beyond the range of the observed values. In cases of extrapolation, the results are always questionable.

The second type of failure occurs when the work practice in the shop changes from the sample used to develop the predictor equation and the sample used to test the model. Work practice includes both the production methods used (e.g., one man vs. two man crew) and the equipment (e.g., a new machine is installed). This type of failure is more serious than one based on a change of work mix, since it requires developing a new database, representing the new work practice.

In the case of shape 1, the predictor equation

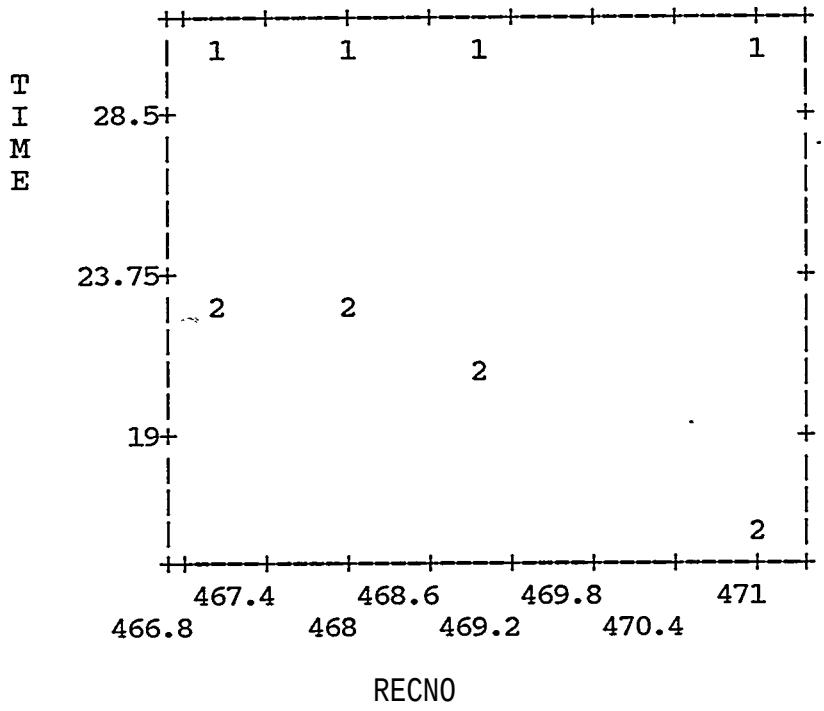
TIME = 1.15 * (X1Y1)

was tested using a sample of 4 records from a later production period. The results of the test are presented in Table 11. Note that the average fabrication time in the test sample Is only 30 minutes, while the average time in the modeling sample is 78 minutes. The sample is clearly at the boundaries of the model's applicability. Also, the sample records are very similar, i.e., they do not represent a cross-section or mix of parts similar to the modeling sample. Despite these potential problems, the prediction error for the sample was only 32%. Clearly, the regression model is an excellent tool for estimating fabrication times.

Table 11. Testing the Regression Model

X1	Y1	TIME observed	TIME predicted	RESIDUAL
6.00	3.25	30.00	22.36	7.64
5.00	4.00	30.00	22.94	7.06
6.00	3.00	30.00	20.64	9.36
3.50	4.00	30.00	16.06	13.94
	TOTAL	120.00	82.00	

Prediction Error = 31.7%



1: TIME observed
2: TIME predicted
$: Multiple occurence

Figure 9. Observed vs Predicted TIME for the Test

6.2.2 Monitoring Model Performance

The problems that arise in testing the model can occur at any point in time, even after the model has successfully passed its initial test. Therefore, application of the predictor equation should incorporate a routine comparison of the predicted versus actual work order direct hours.

The precision with which this comparison is made is not as important as the regularity with which it is made. Even a prediction which is always in error can be useful for shop loading, provided the error is consistent.

BIBILIOGRAPHY

Draper, N. R, and H. Smith, Applied Regression Analysis, 1966, John Wiley and Sons.

Graves, R. J., McGinnis, L. F., and Robinson, R. A., "Standards for Production Planning and Control in Shipyard Shops," Proceedings of IREAPS, San Diego, 1982.

Graves, R. J., McGinnis, L. F., and R. A. Robinson, "Shipyard Production Standards," submitted to Journal of Ship Production.

Hays, W. L, and R. L. Winkler, Statistics: Probability Inference, and Decision, Vol. II, 1970, Holt, Rinehart and Winston, Inc.

McGinnis, L. F. and Graves, R. J., "A Method for Establishing Useful Time Standards for Production Planning and Control in Shipyards," Proceedings of the Symposium on Industrial Engineering Applications in the U. S. Shipbuilding Industry, Institute of Industrial Engineers, 1982.

McGinnis, L. F., and Graves, R. J., "Design of Shipyard Production Standards by Multiple Regression Analysis," in review for journal publication and available from the authors.

Norusis, Marija J., SPSS/PC: For the IBM PC/XT, SPSS Inc., 444 N. Michigan Avenue, Chicago, IL, 1984.

"Scheduling Standards Pilot Project: Summary Report," Ship Producibility Research Program, SNAME Panel SP-8, Bath Iron Works Corporation, 1982.

APPENDIX A

Brief descriptions of the sixteen variables are:

RECNO - Chronological serial number for the line of data
SHAPE - Shape cede
MATL - Material composition, coded as follows:
  01 - Galvanized steel
  02 - Perforated aluminum
  03 - stainless steel
GAUGE - Gauge of material
SEAM - seam type, coded as follows:
  01 - Pittsburgh
  02 - Rivet
  03 - Lock
  04 - Weld
  05 - 3/4" lap
  06 - Spot weld
  07 - Spot weld and rivet
  08 - Lap
  09 - Lockform
TIME - Time, in minutes
X1 - 1st opening height, in inches
Y1 - 1st opening width, in inches
X2 - 2nd opening height, in inches
Y2 - 2nd opening width, in inches
DIAM - Diamter, in inches
ANG - Angle, in degrees
LEN1 - 1st length, in inches
LEN2 - 2nd length, in inches
OFFSET- offset, in inches
NUMPCS- Number of pieces. One piece is assumed, unless an entry appears.
JOINT - Joining method, coded as follows:
  01 - Slip & Drive (S & D)
  02 - S&D + Flange
  03 - Flange RTR Flange
  04 - Flange
  05 - Flange + S&D
  06 - Lock
  07 - Rivet
  08 - Weld
  09 - Flange + S&R
  10 - S&R
  11 - Pittsburgh
  12 - Flange + Rivet
  13 - Bolt
  14 - Spot weld
  15 - Flange + Weld
  16 - Pittsburgh + Rivet
  17 - Pittsburgh + Bolt
  18 - Pittsburgh + S&D
  19 - Spot weld + S&D

Appendix B. Data for Shape 1

```
DATA LIST
/RECNO 1-3 SHAPE 4-6 MATL 7-8 GAUGE 9-11 SEAM 12-13 TIME 14-18 X1 19-25(2)
Y1 26-32(2) X2 33-39(2) Y2 40-46(2) DIAM 47-53(2) ANG 54-56 LEN1 57-63(2)
OFFSET 64-71(2) NUMPCS 72-75 JOINT 76-79.
BEGIN DATA.
  5  1 1 22 2  180  13.00  11.00                                          4
 18  1 1 11 4  210  18.00  16.50  12.00  24.00                            8
 23  1 1 20 2  135  18.00   6.00                 14.25                    2
 24  1 1 20 2  255  15.00   8.00                 14.25                    2
 33  1 1 26 2   45   3.00   4.00                  4.00                    4
 37  1 1 24     75   7.00   3.00                  4.00                    4
 41  1 1 24 5   20  12.00   6.00                 10.00                    7
 46  1 3 26 6   47   2.00   5.00                  3.50                    1
 67  1    22 8  75   9.00   8.00                  9.00     10.00       2  7
 68  1    22 8  45   9.00   8.00                 10.00     11.00       2  7
 89  1 1 18 1  195   5.00  12.00  22.00   6.25
 90  1 1 18 1  285  15.00   5.00  22.00   6.25
 91  1    18 2  630  27.00  21.00                 21.25                   13
 95  1 3 24 6   60   9.00   3.25                  5.00                 2  1
 97  1 3 24 6   41   7.00   3.25                  3.00                    1
 99  1 3 24 6   15   9.00   3.00                  5.00                    1
100  1 3 24 6   45   9.00   3.00                  5.00                   14
101  1 3 24 6   41  10.00   5.00                  6.00                   14
102  1 3 26 6   45   3.00   4.00                  4.00                   14
103  1 3 26 6   90   6.00   4.00                  5.00                   14
104  1 3 26 6   85   6.00   3.00                  3.50                   14
117  1 3 26 6   56   6.00   3.00                  3.00                 2 14
120  1 3 26 6   36   6.00   3.00                  3.00                   14
121  1 3 26 6   60   5.00   1.00                  3.00                   14
122  1 3 22 6   45  14.00   4.00                  4.50                   14
129  1    22 8  15   2.50   6.00                  4.50                    7
133  1 3 22 6   40                                6.00                   14
134  1 3 22 6   24                                4.00                   14
140  1    24 8  20   2.50   6.00                                         7
144  1 3 12 4   37  12.00   6.00                  6.00                 2  8
145  1 3 16 4   45   3.50   6.00                  5.00                    8
146  1 3 16 4   73   4.50   5.00                  6.00                    8
176  1    22 8 180   8.00   4.00                          4.00            7
179  1 1 20 8   60  16.00   5.00   4.50  15.00                            7
191  1 1 20 8  195  16.00   5.00                        17.00              7
193  1 1 24 8   90   4.50   3.00                  3.00    12.00            7
200  1 1 22 1   30  10.00   5.50  12.00   5.00       17  30.00            1
213  1    24 8  45   6.00   3.00                  5.00
274  1 1 26 2   25   5.00   5.00                  5.50                    1
275  1 1 22 2   30   3.25   9.00                  7.00                   10
276  1 1 22 2   50   8.00   5.00                  7.00                 2 10
282  1 1 20 1   30  12.00  10.00  12.00   8.00                            1
284  1 1 26 2   45   3.50   3.50                  4.00                    7
339  1 1 20 2   60   6.00   6.00                 12.88                   12
340  1 1 18 2   90  20.00   7.00                  8.00                    5
341  1 1 20 2   75   5.00   5.00                 12.88                    4
end data.
if (numpcs gt 1) time=time/numpcs.
```